On the Use of Large Language Models for Table Tasks

Yuyang Dong
NEC
Kawasaki, Kanagawa, Japan
dongyuyang@nec.com

Masafumi Oyamada
NEC
Kawasaki, Kanagawa, Japan
oyamada@nec.com

Chuan Xiao
Osaka University, Nagoya University
Suita, Osaka, Japan
chuanx@ist.osaka-u.ac.jp

Haochen Zhang
Osaka University
Suita, Osaka, Japan
chou.koushin@ist.osaka-u.ac.jp

Abstract

The proliferation of large language models (LLMs) has catalyzed a diverse array of applications. This tutorial delves into the application of LLMs for tabular data and targets a variety of table-related tasks, such as table understanding, text-to-SQL conversion, and tabular data preprocessing. It surveys LLM solutions to these tasks in five classes, categorized by their underpinning techniques: prompting, fine-tuning, RAG, agents, and multimodal methods. It discusses how LLMs offer innovative ways to interpret, augment, query, and cleanse tabular data, featuring academic contributions and their practical use in the industrial sector. It emphasizes the versatility and effectiveness of LLMs in handling complex table tasks, showcasing their ability to improve data quality, enhance analytical capabilities, and facilitate more intuitive data interactions. By surveying different approaches, this tutorial highlights the strengths of LLMs in enriching table tasks with more accuracy and usability, setting a foundation for future research and application in data science and AI-driven analytics. Presentation slides for this tutorial will be available at: https://dongyuyang.github.io/tableLLM-tutorial/ .

CCS Concepts

• **Information systems → Specialized information retrieval**; **Information integration**; • **Computing methodologies → Natural language processing**.

Keywords

large language model, tabular data, table tasks

ACM Reference Format:
Yuyang Dong, Masafumi Oyamada, Chuan Xiao, and Haochen Zhang. 2024. On the Use of Large Language Models for Table Tasks. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3627673.3679100

1 Target Audience

This tutorial targets intermediate level of audience, in particular, researchers, developers, and practitioners interested in information retrieval, data science, and artificial intelligence.

We assume that the target audience is generally familiar with basic terms in information retrieval and LLMs, but there is no requirement for prior knowledge of specific algorithms. In addition, this tutorial will provide an introduction to table tasks as well as a primer of LLMs to help the audience understand the techniques used in the tutorial.

2 Relevance to CIKM

The burgeoning interest in table tasks has marked recent years, encompassing a variety of operations such as question answering, fact verification, table enrichment, and relation extraction. These tasks have garnered attention across the IR and AI communities. A notable instance is a tutorial on web tables presented at SIGIR 2019 [61], underscoring the relevance of these topics to the IR community.

Our tutorial extends this discourse by covering an expansive array of table tasks and exploring LLM methodologies for addressing these challenges. Compared with recent surveys on the same topic [14, 38], our tutorial delves into retrieval tasks and solutions, with a particular emphasis on RAG approaches to table tasks. It illustrates how cutting-edge IR techniques are employed to enhance the performance of LLMs in handling table tasks, thereby providing tangible insights for IR researchers and practitioners.

Furthermore, to underscore the applicability and practical relevance of the methods surveyed, our tutorial incorporates case studies from the industrial sector. This includes an in-depth look at the processing of in-document tables, such as those found within web pages and PDF documents, highlighting their significance and the complexities involved in their retrieval and interpretation.

Given its focus on the use of LLMs in mining and retrieval tasks, this tutorial will not only provide an overview of current methodologies and their applications but also foster a deeper understanding of the potential for future innovations in AI, IR, and DM. We are confident that it will be a valuable contribution to CIKM, appealing to academic researchers and industry professionals in these fields.

3 Benefits

Overall, the tutorial is designed to not only educate but also inspire participants, fostering a community that is well-informed, ethically conscious, and capable of leveraging AI technologies to address complex problems in IR and improve societal outcomes. Specifically, the

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679100

objectives of this tutorial are outlined as follows: (1) **Empowering Professionals**: The audience will gain a comprehensive understanding of the use of LLMs in handling table tasks, grasp the basic concepts of state-of-the-art solutions, become familiar with recent research trends, and learn how to select appropriate solutions for various application scenarios, thereby driving innovation and efficiency in their respective fields. (2) **Bridging Knowledge Gaps**: The tutorial aims to demystify the complexities of LLMs for a broader audience, reducing the knowledge gap between cutting-edge research and industry practitioners, hence fostering a more informed and technologically adept society. (3) **Promoting Ethical Use of AI**: Through discussions on the ethical implications and responsible use of LLMs in IR, we encourage participants to adopt practices that prioritize fairness, privacy, and transparency, contributing to the development of more ethical AI solutions. (4) **Stimulating Research and Collaboration**: By showcasing the latest trends and challenges in the field, the tutorial aims to serve as a catalyst for new research initiatives and collaborations across academia and industry.

4 Outline

This is a 3-hour lecture-style tutorial, divided into seven parts, each concluding with a Q&A session to encourage audience interaction and clarification of concepts. A coffee break is scheduled after the first three parts to allow participants to refresh and network. The detailed schedule is presented in Table 1.

4.1 Introduction & Preliminaries

This part underscores the importance of table tasks and lays the groundwork with fundamental concepts. We start by defining types of tabular data, including relational tables, spreadsheets, and hierarchical tables. Subsequently, we categorize table tasks into understanding, manipulation, augmentation, preprocessing, and text-to-SQL, followed by introducing evaluation metrics for these tasks and reviewing publicly available benchmarks [6, 18, 25, 27, 57]. To review solutions to table tasks, we begin with a summary of approaches prior to the advent of LLMs, e.g., table representation learning (TRL) methods [9, 19, 22, 56]. We then introduce the basics of LLMs, discussing the advantages and limitations of using LLMs for table tasks alongside key techniques for crafting LLM solutions.

4.2 Prompting

This part first explores the transformation of raw tabular data into prompts suitable for LLM input, including methods to serialize tables, rows, or columns into text. We delve into common prompt engineering techniques such as few-shot [5], zero-shot prompting [24], chain-of-thought [51], and batch prompting [7], along with table-specific prompting strategies [15]. Representative methods for table understanding [4, 23, 26], table augmentation [46], data preprocessing [39, 60], and text-to-SQL [41] are surveyed, including industrial examples and empirical results [45].

4.3 Fine-Tuning

This part highlights the distinctiveness of LLMs from previous models, emphasizing their capability for instruction tuning [62]. We begin with criteria for base model selection and data preparation for instruction tuning, followed by instruction data construction for table tasks. We review instruction-tuned models like Table-GPT [34], based on GPT-3.5 or ChatGPT, and TableLlama [63], based on Llama-7B, comparing them to models like UnifiedSKG [55] from the transition period from pre-trained language models (PLMs) to LLMs. Jellyfish [59], an instruction-tuned model for tabular data preprocessing, serves as a case study to illustrate techniques in this part.

Table 1: Tutorial outline (3 hours + break).

Part 1 (30 minutes): Introduction & preliminaries
1.1 Tabular data
1.2 Table tasks
1.3 Benchmarks
1.4 Summary of approaches prior to the rise of LLMs
1.5 LLM preliminaries
1.6 Key techniques for the use of LLMs in table tasks
1.7 Q&A session
Part 2 (30 minutes): Prompting
2.1 Transformation of tabular data to prompts
2.2 Prompting techniques
2.3 Representative solutions and examples
2.4 Q&A session
Part 3 (30 minutes): Fine-tuning
3.1 Base model selection
3.2 Data preparation
3.3 Construction of instruction data
3.4 Fine-tuning techniques
3.5 Representative solutions and examples
3.6 Q&A session
Break
Part 4 (35 minutes): Retrieval-augmented generation (RAG)
4.1 Basic RAG techniques
4.2 Advanced RAG techniques
4.3 RAG for table tasks
4.4 Representative solutions and examples
4.5 Q&A session
Part 5 (20 minutes): LLM agents
5.1 LLM agent preliminaries
5.2 LLM agents for table tasks
5.3 Representative solutions and examples
5.4 Q&A session
Part 6 (20 minutes): Vision-language models (VLMs)
6.1 VLM preliminaries
6.2 Comparison of VLM and LLM methods for table tasks
Part 7 (15 minutes): Conclusions
7.1 Summary of state-of-the-art
7.2 Issues and future research directions
7.3 Societal impacts
7.4 Q&A session

4.4 Retrieval-Augmented Generation

This part introduces RAG, giving LLMs access to external information to enhance generation performance [30, 33]. Starting with basic techniques like chunking and embedding-based retrieval, we explore advanced RAG methods particularly beneficial for table tasks. After RAG methods developed for tables [35, 36] are reviewed, benchmarking RAG on tabular data [28] and observations on table retrieval's effectiveness without the necessity for table-specific models (e.g., TaPas [19]) and the sufficiency of text models (e.g., BERT [10]) for embedding are discussed [27].

4.5 LLM Agents

The prevalence of LLMs is accompanied by the rise of LLM agents [50] as tools for engineering [20] or simulation [52]. We outline the components of LLM agents, including action, memory, planning, and tool use, and how table task processing benefits from LLM agents, showcasing methods like SheetCopilot [32] and ReAcTable [64].

4.6 Vision-Language Models

Focusing on vision-language models (VLMs), this part examines their application in processing tables in scanned documents without OCR. Beginning with the preliminaries of VLMs, we compare VLMs and LLMs in table tasks, summarizing empirical findings [8].

4.7 Conclusions

We summarize the state-of-the-art LLM methods for table tasks, and then discuss future research directions and potential societal impacts. The tutorial ends with a comprehensive Q&A session.

5 Related Tutorials

This tutorial has not been previously presented.

In the realm of tabular data, several related tutorials have previously been presented at various conferences, including: (1) SIGIR 2019 (Paris, Jul. 21 – 25, 2019): "Web table extraction, retrieval and augmentation" [61]; (2) KDD 2021 (Singapore, Aug. 14 – 18, 2021): "From Tables to Knowledge: Recent Advances in Table Understanding" [42]; (3) VLDB 2022 (Sydney, Sep. 5 – 9, 2022) & SIGMOD 2023 (Seattle, Jul. 18 – 23, 2023): "Transformers for Tabular Data Representation: A tutorial on Models and Applications" [2] and "Models and Practice of Neural Table Representations" [21] as a new edition; (4) ICDE 2023 (Anaheim, Apr. 3 – 7, 2023) & WWW 2023 (Austin, Apr. 30 – May 4, 2023): "Graph Neural Networks for Tabular Data Learning" [31]. These tutorials focused on methodologies that predate the LLM era, especially on PLMs such as BERT [10] and RoBERTa [37]. In contrast, the LLM methods surveyed in our tutorial are designed to function in a prompted manner, offering downstream applications a seamless end-to-end solution. This distinction underlines the unique contribution and perspective our tutorial brings to the field, emphasizing the innovative use of LLMs for direct application in table tasks. In addition, our tutorial covers VLM methods, featuring understanding diverse table images and textual instructions.

6 Brief Biographies

- **Yuyang Dong**is a Principal Researcher at NEC. He earned his Ph.D. degree from the University of Tsukuba in 2019. He specializes in tabular data searching and NLP, with his expertise in **tabular data processing** evidenced by publications in prestigious venues such as VLDB 2023 [13], SIGIR 2022 [11], and ICDE 2021 [12].

  Dong leads the project **Jellyfish** [59], a leading-edge LLM for tabular data processing that has garnered thousands of monthly downloads on Hugging Face [58]. He plays a pivotal role in the development of the NEC data enrichment service, an AI-driven solution aimed at enhancing data quantity and quality. Additionally, he is a key contributor to NEC **cotomi-core**, a suite of robust, self-developed LLMs in both Japanese and English, underpinning NEC's Generative AI Service.
- **Masafumi Oyamada**is the Chief Scientist and Director of Generative AI Foundations Research at NEC. He was awarded his Ph.D. degree from the University of Tsukuba in 2018. At NEC, he spearheads research and development efforts in the domain of LLMs, including the creation of **cotomi-core**, a series of high-performance and efficient LLMs. His interdisciplinary work has bridged the gap between **tabular data and machine learning**. His contributions encompass probabilistic modeling for entity-relationship tabular data (ICDM 2017 [40]), probabilistic methods for labeling tabular data (AAAI 2019 [48]), and tabular data discovery (ICDE 2021 [12], SIGIR 2022 [11], and VLDB 2023 [13]).

  More recently, his focus has expanded to the exploration of **LLMs for structured data**, leading to advances in knowledge extraction from LLMs (EMNLP 2021 [47]), enhancing entity matching via question-answering models (PAKDD 2023 [17]), developing dense retrieval models tailored for tabular data (VLDB 2023 [13]), formulating organizational LLMs (BigData 2023 [3]), and conducting bias analysis for RAG (EMNLP 2023 [1]).
- **Chuan Xiao**is an Associate Professor at Osaka University and a Guest Associate Professor at Nagoya University. He completed his Ph.D. at the University of New South Wales in 2010. His research interests span data preprocessing, computational social science, and NLP. He has been actively engaged in research related to prompting [60] and fine-tuning [59] LLMs for data preprocessing, leading to the development of **Jellyfish** [59] – a cutting-edge LLM designed specifically for this purpose. His work on **tabular data discovery** has been recognized at prestigious conferences like VLDB 2023 [13] and ICDE 2021 [12]. He introduced the concept of **smart agent-based modeling** [52], a computer simulation approach that utilizes LLM agents for exploring social sciences (e.g., economy [16], behavioral science [53], and urban computing [49]).

  With over 15 years of research experience in similarity search, Xiao has published 20+ related papers at top-tier conferences (ICML, SIGMOD, VLDB, WWW, etc.). He is the principal developer of the PPJoin algorithm [54], a renowned near-duplicate detection algorithm featured in the data mining textbook "Mining Massive Datasets" co-authored by Jeffrey D. Ullman [29]. Furthermore, he has conducted **tutorials** at KDD 2021 [44] and VLDB 2020 [43].
- **Haochen Zhang**is currently pursuing his Master's degree at Osaka University. He obtained his Bachelor's degree from Osaka City University in 2023. His research areas of interest encompass data mining, data preprocessing, and NLP. During his internship at NEC, he focused on exploring prompting and fine-tuning LLMs for data preprocessing ([60]). He has contributed to the development of **Jellyfish** [59] – a leading-edge LLM specifically designed for data preprocessing. Additionally, he serves as a teaching assistant for the big data engineering course at Osaka University.

Acknowledgments

This work is supported by NEC Corporation, JSPS Kakenhi 23K17456, 23K25157, 23K28096, and CREST JPMJCR22M2.

References

[1] K. Akimoto, K. Takeoka, and M. Oyamada. Context quality matters in training fusion-in-decoder for extractive open-domain question answering. *arXiv preprint*

arXiv:2403.14197, 2024.

[2] G. Badaro and P. Papotti. Transformers for tabular data representation: A tutorial on models and applications. *PVLDB*, 15(12):3746–3749, 2022.

[3] K. Boros and M. Oyamada. Towards large language model organization: A case study on abstractive summarization. In *BigData*, pages 6109–6112, 2023.

[4] A. Brinkmann, R. Shraga, and C. Bizer. Product attribute value extraction using large language models. *arXiv preprint arXiv:2310.12537*, 2023.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020.

[6] W. Chen, H. Wang, J. Chen, Y. Zhang, H. Wang, S. Li, X. Zhou, and W. Y. Wang. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*, 2019.

[7] Z. Cheng, J. Kasai, and T. Yu. Batch prompting: Efficient inference with large language model APIs. *arXiv preprint arXiv:2301.08721*, 2023.

[8] N. Deng, Z. Sun, R. He, A. Sikka, Y. Chen, L. Ma, Y. Zhang, and R. Mihalcea. Tables as images? exploring the strengths and limitations of llms on multimodal representations of tabular data. *arXiv preprint arXiv:2402.12424*, 2024.

[9] X. Deng, H. Sun, A. Lees, Y. Wu, and C. Yu. TURL: Table understanding through representation learning. *ACM SIGMOD Record*, 51(1):33–40, 2022.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[11] Y. Dong and M. Oyamada. Table enrichment system for machine learning. In *SIGIR*, pages 3267–3271, 2022.

[12] Y. Dong, K. Takeoka, C. Xiao, and M. Oyamada. Efficient joinable table discovery in data lakes: A high-dimensional similarity-based approach. In *ICDE*, pages 456–467, 2021.

[13] Y. Dong, C. Xiao, T. Nozawa, M. Enomoto, and M. Oyamada. Deepjoin: Joinable table discovery with pre-trained language models. *PVLDB*, 16(10):2458–2470, 2023.

[14] X. Fang, W. Xu, F. A. Tan, J. Zhang, Z. Hu, Y. Qi, S. Nickleach, D. Socolinsky, S. Sengamedu, and C. Faloutsos. Large language models on tabular data–a survey. *arXiv preprint arXiv:2402.17944*, 2024.

[15] Google Research. Chain-of-table: Evolving tables in the reasoning chain for table understanding, 2024.

[16] X. Han, Z. Wu, and C. Xiao. "guinea pig trials" utilizing gpt: A novel smart agent-based modeling approach for studying firm competition and collusion. *arXiv preprint arXiv:2308.10974*, 2023.

[17] S. Hayashi, Y. Dong, and M. Oyamada. Qa-matcher: Unsupervised entity matching using a question answering model. In *PAKDD*, pages 174–185, 2023.

[18] X. He, M. Zhou, M. Zhou, J. Xu, X. Lv, T. Li, Y. Shao, S. Han, Z. Yuan, and D. Zhang. Anameta: A table understanding dataset of field metadata knowledge shared by multi-dimensional data analysis tasks. *arXiv preprint arXiv:2209.00946*, 2022.

[19] J. Herzig, P. K. Nowak, T. Müller, F. Piccinno, and J. M. Eisenschlos. TaPas: Weakly supervised table parsing via pre-training. *arXiv preprint arXiv:2004.02349*, 2020.

[20] S. Hong, X. Zheng, J. Chen, Y. Cheng, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou, C. Ran, et al. MetaGPT: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2023.

[21] M. Hulsebos, X. Deng, H. Sun, and P. Papotti. Models and practice of neural table representations. In *SIGMOD*, pages 83–89, 2023.

[22] H. Iida, D. Thai, V. Manjunatha, and M. Iyyer. Tabbie: Pretrained representations of tabular data. *arXiv preprint arXiv:2105.02584*, 2021.

[23] J. Jiang, K. Zhou, Z. Dong, K. Ye, W. X. Zhao, and J.-R. Wen. Structgpt: A general framework for large language model to reason over structured data. *arXiv preprint arXiv:2305.09645*, 2023.

[24] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022.

[25] P. Konda, S. Das, A. Doan, A. Ardalan, J. R. Ballard, H. Li, F. Panahi, H. Zhang, J. Naughton, S. Prasad, et al. Magellan: toward building entity matching management systems over data science stacks. *PVLDB*, 9(13):1581–1584, 2016.

[26] K. Korini and C. Bizer. Column type annotation using ChatGPT. *arXiv preprint arXiv:2306.00745*, 2023.

[27] S. Kweon, Y. Kwon, S. Cho, Y. Jo, and E. Choi. Open-wikitable: Dataset for open domain question answering with complex reasoning over table. *arXiv preprint arXiv:2305.07288*, 2023.

[28] LangChain. Benchmarking RAG on tables, 2023.

[29] J. Leskovec, A. Rajaraman, and J. D. Ullman. *Mining of massive data sets*. Cambridge university press, 2020.

[30] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *NeurIPS*, 33:9459–9474, 2020.

[31] C.-T. Li, Y.-C. Tsai, and J. C. Liao. Graph neural networks for tabular data learning. In *ICDE*, pages 3589–3592, 2023.

[32] H. Li, J. Su, Y. Chen, Q. Li, and Z.-X. Zhang. Sheetcopilot: Bringing software productivity to the next level through large language models. *NeurIPS*, 36, 2024.

[33] H. Li, Y. Su, D. Cai, Y. Wang, and L. Liu. A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110*, 2022.

[34] P. Li, Y. He, D. Yashar, W. Cui, S. Ge, H. Zhang, D. R. Fainman, D. Zhang, and S. Chaudhuri. Table-GPT: Table-tuned GPT for diverse table tasks. *arXiv preprint arXiv:2310.09263*, 2023.

[35] W. Lin, R. Blloshmi, B. Byrne, A. de Gispert, and G. Iglesias. An inner table retriever for robust table question answering. In *ACL*, pages 9909–9926, 2023.

[36] W. Lin, R. Blloshmi, B. Byrne, A. de Gispert, and G. Iglesias. Li-rage: Late interaction retrieval augmented generation with explicit signals for open-domain table question answering. In *ACL*, pages 1557–1566, 2023.

[37] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[38] W. Lu, J. Zhang, J. Zhang, and Y. Chen. Large language model for table processing: A survey. *arXiv preprint arXiv:2402.05121*, 2024.

[39] A. Narayan, I. Chami, L. Orr, and C. Ré. Can foundation models wrangle your data? *PVLDB*, 16(4):738–746, 2022.

[40] M. Oyamada and S. Nakadai. Relational mixture of experts: Explainable demographics prediction with behavioral data. In *ICDM*, pages 357–366, 2017.

[41] M. Pourreza and D. Rafiei. DIN-SQL: Decomposed in-context learning of text-to-SQL with self-correction. *NeurIPS*, 36, 2023.

[42] J. Pujara, P. Szekely, H. Sun, and M. Chen. From tables to knowledge: Recent advances in table understanding. In *KDD*, pages 4060–4061, 2021.

[43] J. Qin, W. Wang, C. Xiao, and Y. Zhang. Similarity query processing for high-dimensional data. *PVLDB*, 13(12):3437–3440, 2020.

[44] J. Qin, W. Wang, C. Xiao, Y. Zhang, and Y. Wang. High-dimensional similarity query processing for data science. In *KDD*, pages 4062–4063, 2021.

[45] Y. Sui, M. Zhou, M. Zhou, S. Han, and D. Zhang. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *WSDM*, pages 645–654, 2024.

[46] Y. Sui, J. Zou, M. Zhou, X. He, L. Du, S. Han, and D. Zhang. TAP4LLM: Table provider on sampling, augmenting, and packing semi-structured data for large language model reasoning. *arXiv preprint arXiv:2312.09039*, 2023.

[47] K. Takeoka, K. Akimoto, and M. Oyamada. Low-resource taxonomy enrichment with pretrained language models. In *EMNLP*, pages 2747–2758, 2021.

[48] K. Takeoka, M. Oyamada, S. Nakadai, and T. Okadome. Meimei: An efficient probabilistic approach for semantically annotating tables. In *AAAI*, volume 33, pages 281–288, 2019.

[49] J. Wang, R. Jiang, C. Yang, Z. Wu, M. Onizuka, R. Shibasaki, and C. Xiao. Large language models as urban residents: An llm agent framework for personal mobility generation. *arXiv preprint arXiv:2402.14744*, 2024.

[50] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, et al. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*, 2023.

[51] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, and D. Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[52] Z. Wu, R. Peng, X. Han, S. Zheng, Y. Zhang, and C. Xiao. Smart agent-based modeling: On the use of large language models in computer simulations. *arXiv preprint arXiv:2311.06330*, 2023.

[53] Z. Wu, S. Zheng, Q. Liu, X. Han, B. I. Kwon, M. Onizuka, S. Tang, R. Peng, and C. Xiao. Shall we talk: Exploring spontaneous collaborations of competing llm agents. *arXiv preprint arXiv:2402.12327*, 2024.

[54] C. Xiao, W. Wang, X. Lin, J. X. Yu, and G. Wang. Efficient similarity joins for near-duplicate detection. *ACM Transactions on Database Systems*, 36(3):1–41, 2011.

[55] T. Xie, C. H. Wu, P. Shi, R. Zhong, T. Scholak, M. Yasunaga, C.-S. Wu, M. Zhong, P. Yin, S. I. Wang, et al. UnifiedSKG: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. *arXiv preprint arXiv:2201.05966*, 2022.

[56] P. Yin, G. Neubig, W.-t. Yih, and S. Riedel. TaBERT: Pretraining for joint understanding of textual and tabular data. *arXiv preprint arXiv:2005.08314*, 2020.

[57] T. Yu, R. Zhang, K. Yang, M. Yasunaga, D. Wang, Z. Li, J. Ma, I. Li, Q. Yao, S. Roman, et al. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. *arXiv preprint arXiv:1809.08887*, 2018.

[58] H. Zhang, Y. Dong, C. Xiao, and M. Oyamada. Jellyfish. https://huggingface.co/NECOUDBFM/Jellyfish, 2023.

[59] H. Zhang, Y. Dong, C. Xiao, and M. Oyamada. Jellyfish: A large language model for data preprocessing. *arXiv preprint arXiv:2312.01678*, 2023.

[60] H. Zhang, Y. Dong, C. Xiao, and M. Oyamada. Large language models as data preprocessors. *arXiv preprint arXiv:2308.16361*, 2023.

[61] S. Zhang and K. Balog. Web table extraction, retrieval and augmentation. In *SIGIR*, pages 1409–1410, 2019.

[62] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.

[63] T. Zhang, X. Yue, Y. Li, and H. Sun. Tablellama: Towards open large generalist models for tables. *arXiv preprint arXiv:2311.09206*, 2023.

[64] Y. Zhang, J. Henkel, A. Floratou, J. Cahoon, S. Deep, and J. M. Patel. Reactable: Enhancing react for table question answering. *arXiv preprint arXiv:2310.00815*, 2023.